Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 404 – 411

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

# ATSSI: Abstractive Text Summarization using Sentiment Infusion

Rupal Bhargava*, Yashvardhan Sharma and Gargi Sharma

*Birla Institute of Technology & Science, Pilani, Pilani Campus 333 031, India*

**Abstract**

Text Summarization is condensing of text such that, redundant data are removed and important information is extracted and represented in the shortest way possible. With the explosion of the abundant data present on social media, it has become important to analyze this text for seeking information and use it for the advantage of various applications and people. From past few years, this task of automatic summarization has stirred the interest among communities of Natural Language Processing and Text Mining, especially when it comes to opinion summarization. Opinions play a pivotal role in decision making in the society. Other's opinions and suggestions are the base for an individual or a company while making decisions. In this paper, we propose a graph based technique that generates summaries of redundant opinions and uses sentiment analysis to combine the statements. The summaries thus generated are abstraction based summaries and are well formed to convey the gist of the text.


Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

*Keywords:* Abstractive Summarization; Condensed Text; Data Redundancy; Sentiment Analysis; Summary; Text Summarization.

## 1. Introduction

There is a large amount of data on the web which expresses the same opinion over and over again. Summarization of dispensable content, thus is a necessity. While viewing multi-document summaries or the summaries of highly redundant text, extractive summarization would not be of any help as the extractive summaries would be very verbose and biased. Also, the sentences tend to be longer, hence non-essential parts of the sentence also get included. Relevant information is spread across the document and this can't be captured in the extractive summaries. Extractive summaries also face the problem of "dangling" anaphora, implying that sentences that contain pronouns lose meaning when extracted out of context, the resolution of which is presented in Steinberger J. *et al.*[2].

While there has been a lot of work done in the field of extraction based summarization, abstraction based summarization is difficult because of the simple reason that while the computers can statistically select the most important sentence from the text, it is difficult for them to combine important sentences and generate a coherent and concise synopsis. Demand for high quality summary is on the rise whether it is regarding summarization of textual content (for example books etc.) or multimedia content like video transcripts etc. (Ding, Duo, *et al.*[3]).

It has been demonstrated that abstractive summaries perform well than extractive summaries (Carenini, G. *et al.*[4]) whenever documents with a lot of redundant content (e.g. Product reviews, blogs and news articles, etc.). This is

*Corresponding author. Tel.: +91 -9829474312.
*E-mail address:* rupal.bhargava@pilani.bits_ pilani.ac.in


Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.088

because abstractive summaries are compact and present the useful information and are not verbose. But, generating abstract summary is a tougher task than generation of extract summary. Also, it should be noted that single document summarization is somewhat not quite the same as multi document synopsis, since single documents contain lesser data. Thus, a more efficient strategy is required to generate abstractive summaries in case of single documents.

In this paper, method is proposed for compressing and merging information based on word graphs, and then summaries are generated from the resulting sentences. The method assumes no domain knowledge and leverages redundancy in the text. The results show that the summaries generated are agreeable to human compendium and are concise and well formed.

The paper has been split into three sections. In the related works section, previous work on recent abstractive summarization techniques is explained and novelty our approach is stressed with respect to preexisting frameworks for graph based abstractive summarization. In the methodology section, we describe the algorithm that has been used for summarization. In the Results and discussion Section, we present the results of our algorithm and a detailed analysis of the results. In the Conclusion and future work Section, we provide an insight into the possible areas that can be explored in terms of summarization.

## 2. Related Work

Abstractive techniques in text summarization include rule based approach (Genest, P. E. *et al.*[1]), sentence compression (Knight, K. *et al.*[5], Zajic, D. *et al.*[6], Clarke, J. *et al.*[7]), merging sentence based on their semantics (Liu F. *et al.*[8], Wang D. *et al.*[9]), etc. Graph based algorithms, in particular has been proven to work well on both summarizing texts containing lots of redundant data (Ganesan, K. *et al.*[10], Lloret E., *et al.*[11]), etc.

Sankarasubramaniam Y. *et al.*[23] leverage wikipedia in addition to graph based algorithms to generate extractive summaries. They first map all the sentences to corresponding Wikipedia topic and thus a bipartite graph is obtained where one of vertices represent the wikipedia topics and the other set represent the sentences in the document.[23] then uses an iterative ranking algorithm to find the best candidate sentences in the document.[23] also introduces incremental summarization wherein longer summaries are generated in real-time by simply adding sentences to shorter summaries. Since the summaries generated are extractive, the precision is less when compared to the results of techniques that generate abstractive summaries.

Liu F. *et al.*[8] use the advances in the semantic representation of the text in the form of Abstract Meaning Representation graphs to form summaries. The summarization framework consists of parsing input sentences to form individual AMR graphs, combining the individual AMR graphs to form a summary AMR graph and then generating text from the summary graph. The individual graphs are converted to summary graph using a perceptron model prediction algorithm which predicts with a high accuracy the subgraph that has to be selected for summary generation.

Ganesan K. *et al.*[10] describe an approach that used directed graphs that use the original sentence word order to generate abstractive summaries. Their technique leverages the graphical form of the input text to reduce redundancy. If their algorithm finds two sentences that are collapsible, they use the connectors already present in one of the sentences to be used as the connector for the collapsed sentence. While this technique is effective this still has a drawback since there might be two sentences which are capable of being fused together, but can't be fused because of the absence of a pre-existing connector. Our approach does not face this drawback since we use sentiment analysis to overcome this issue.

Lloret E., *et al.*[11] describes a technique in which they have built a directed weighted word graph where each word text represents a node in the graph and the edge contains the adjacency relation between the words. The weight of the edge is determined by using a combination of their pagerank value and the frequency of the words. To determine important sentences, the first node consists of the first ten words with highest TF-IDF score. Sentence correctness are ensured using the basic rules of grammar like the length a sentence should be greater than 3 words, a sentence must contain a verb and should not end in an article or conjunction. A huge flaw with this methodology is that a lot of important information is lost because of the impositions of grammar on the sentences and the policy of selecting the ten words with highest TF-IDF scores. Furthermore, a lot of redundant sentences will still be present in the summary because the TF-IDF scores will give more importance to them. Our methodology does not face the deficiency that[11] faces because it incorporate the redundancies in our graph structure itself.